Asian Journal of Applied Science and Technology (AJAST)
Volume 7, Issue 1, Pages 132-138, January-March 2023

Artificial Intelligent Decoding of Rare Words in Natural Language Translation Using Lexical Level Context

D. Samson[1*], Dr. Ignatius A Herman[2], Dr. Ibrahim Olanya Bwalya[3] & Dr. Kitraj Penchinkilas[4]

[1]Research Scholar, Gideon Robert University, Lusaka, Zambia. [2]Director-DMI Group of Institutions, Africa. [3]Professor, Gideon Robert University, Lusaka, Zambia. [4]Department of Computer Science, Gideon Robert University, Lusaka, Zambia.
*Corresponding Author (D. Samson) - swansamson@gmail.com**

DOI: https://doi.org/10.38177/ajast.2023.7112

Article Received: 29 January 2023 *Article Accepted: 28 February 2023* *Article Published: 30 March 2023*

ABSTRACT

Artificial intelligence based Machine Translation is a Natural Language Processing, attaining significant attention to fully automate the system that can translate source language content into the target languages. The proposed method is a Neural Machine Translation, end-to-end system, with the lexical level context for converting German sentences to English sentences. The Transformer based Machine Translation is used for translation for handling the occurrence of rare words in the lexical level context. The FSR group rare words together and represent sentence level context as Latent Topic Representation. WMT En-De bilingual parallel corpus is used for translation handling the Out of Vocabulary words using clustering of <UNK> tags. In the existing method there is a mismatch of translation but the proposed system is more superior due to the sentence context inclusion. The model performance is enhanced with hyper parameter optimization obtaining a BLEU score with a better translation of source to target language. Finally minimizing the TER score to attain a better translation rate.

Keywords: Artificial intelligence; Natural language translation; Natural language processing; Machine learning.

░ Objective

(a) To build an end-to-end solution for the language translation using NMT;

(b) To detect the source sentence and target sentence rare words using FSR with hierarchical clustering and perform embedding;

(c) To design and train CNN to estimate the sentence context as LTR;

(d) To obtain the internal vector representation by training the encoder model using source embedding [Seq-to-Vec representation];

(e) To predict the target word by training the decoder model using target embedding [Vec-to-Seq representation].

░ 1. Introduction

As in the current scenario there exist several languages in the world. All the languages can’t be understood, so there is a need for a language translator between the source and target language. Natural Language Translation, which is a NLP task, is the process of converting source language into the target language using a Seq-Seq Encoder - Decoder model [1]. A machine translator is a NLP task which solves the problem of taking the input sentence of the source language and given to the Neural Machine Translator (NMT) for processing the language and gives the output of the target language [2]. Neural machine translation (NMT) uses a model that considers the context of the entire sentence to produce an accurate translation. In this method, an artificial neural network is used to calculate the likelihood of a given word sequence. Using this approach, the translation process may proceed without a hitch [3]. The modules required to build an end-to-end Encoder Decoder system for translation is described here. From the figure 1 the high-level block diagram where the dataset of bilingual parallel English-German corpus is preprocessed to get a clean sentence by removing special characters and punctuation marks. The cleaned dataset is splitted into separate files of German sentences and English sentences. The words which are OOV are given to the FSR most commonly related together under a universal <UNK> tag [4].

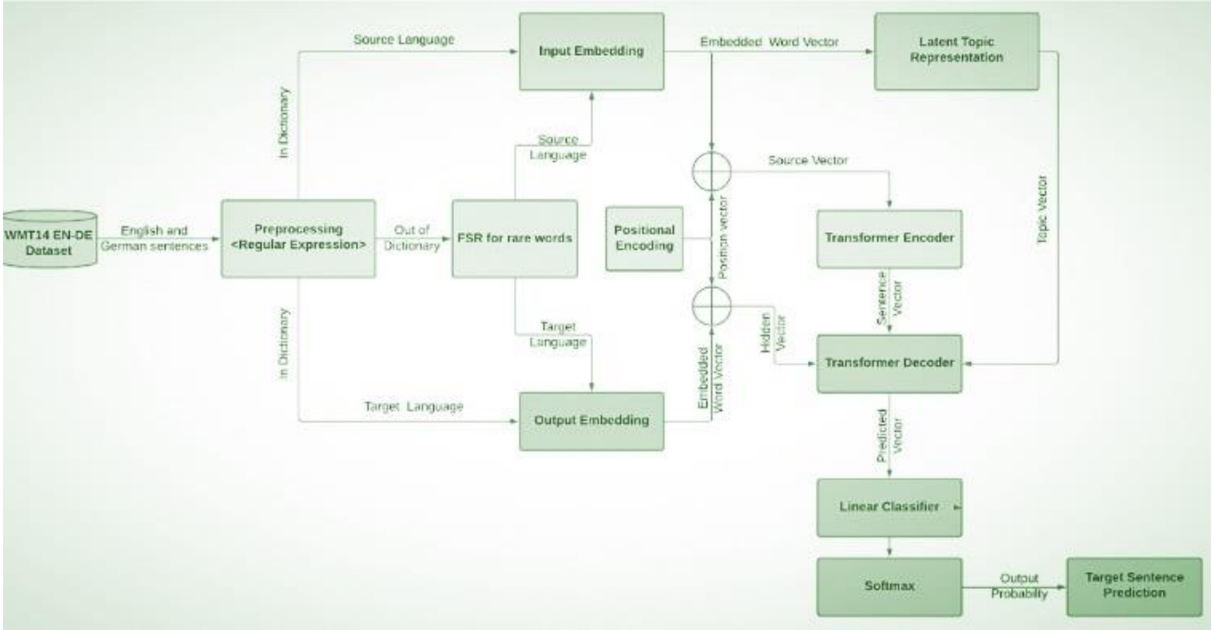

Figure 1. High level block diagram

The embedding module converts the sentence sequence into its corresponding vector representation. The positional encoding module uses sinusoidal functions to indicate the location of the words in the sentence at both source language and target side language [5]. The context of the sentence is handled as Latent Topic Representation which uses CNN to extract the context of the sentence [6-12]. The transformer encoder and decoder which uses a self-attention mechanism to translate. Finally, the decoded vector is passed to the linear classifier and softmax functions. Finally, the google translator is integrated into our NMT system to check the correctness of the predicted sentence using the sentence prediction module.

2. Proposed Method

2.1. Dataset description

Dataset Name - WMT’14

Accessibility - Free

Modality - Text sentences

2.2. Description

Train (4.5M sentence pairs): [train.en] [train.de]

Test:[newstest2012.en],[newstest2012.de],[newstest2013.en],[newstest2013.de],

[newstest2014.en],[newstest2014.de],[newstest2015. en],[newstest2015.de]

Vocabularies (top 50K frequent words): [vocab.50K.en] [vocab.50K.de]

Dictionary (extracted from alignment data): [dict.en-de]

2.3. Experimental Setup

The proposed model is a Transformer based Neural Machine Translation with Self Attention Mechanism which is added at both the Encoder and Decoder layer for translating the source language sentence (German) to the target language sentence (English). The objective is to build an end to end system for Language Translation using NMT. From the dataset 50K sentence pairs are taken and split in the ratio of 80:20 for the training and testing purpose respectively.

A. Pre-Processing

Data preprocessing is a vital phase in the data mining process. The preprocessing module takes care of the representation and quality of data firs t and foremost before running any analysis. The translation corpus includes data loading, data cleaning and data splitting.

B. Data Loading

The dataset is loaded in google drive which is zipped file. The path to the file should be mentioned for downloading the data and unzipping 'En-De.zip' file and extracting Bilingual corpus Deu.txt file.

C. Data Cleaning

The file contains punctuations, uppercase, lowercase and special German characters which is not of much importance for translation purposes so it can be removed using the Regular Expression. The file is converted from Unicode to ASCII characters for each word in a sentence. Finally <start> and <end> tags are added for every sentence which indicates the start and end of Encoder and Decoder respectively.

D. Data Splitting

The cleaned data of the corpus is stripped to break the bilingual corpus line by line. Every line contains the English and its corresponding German sentence. The next step is to split the bilingual sentences into English and German sentences separately. German sentences are given to the encoder and English sentences are given to the decoder. The training set with sentence length less than or equal to 50 is taken for consideration at both source and target files. Padding is done to make all sentences of the same length so that encoding and decoding becomes easier.

E. Word Embedding

INPUT: English Sentence (Input Embedding)

German Sentences (Output Embedding) OUTPUT: Word vector representation The embedding module is responsible for SeqtoVec conversion which specifically uses the Word2Vec model for this purpose. The layers of the Neural Network are dense and fully connected to attain better translation prediction.

The input layer node consists of every word in a sentence which is a vector and based on the predefined weights from CBOW model is used of size VxN for each word with C words in a sentence. The hidden layer is the weighted sum of input nodes and its corresponding weights with a linear activation function and output layer gives the vectors of a sentence which uses sigmoid activation function.

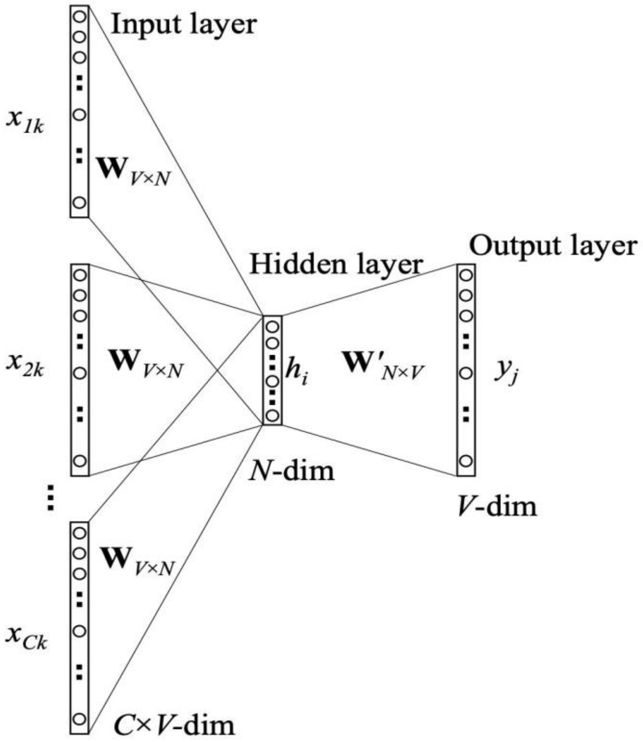

Figure 2. Layer Representation of Word Embedding

The results obtained from the Neural Machine Translation for German to English Conversion is discussed in this paper. The preprocessing, word tokenizer and the translation sentences are added as the snapshots. The Google Translate module is integrated for the checking correct translation of the German sentences to its corresponding English sentences. The Model is trained by limiting the corpus to 50000 sentence examples.

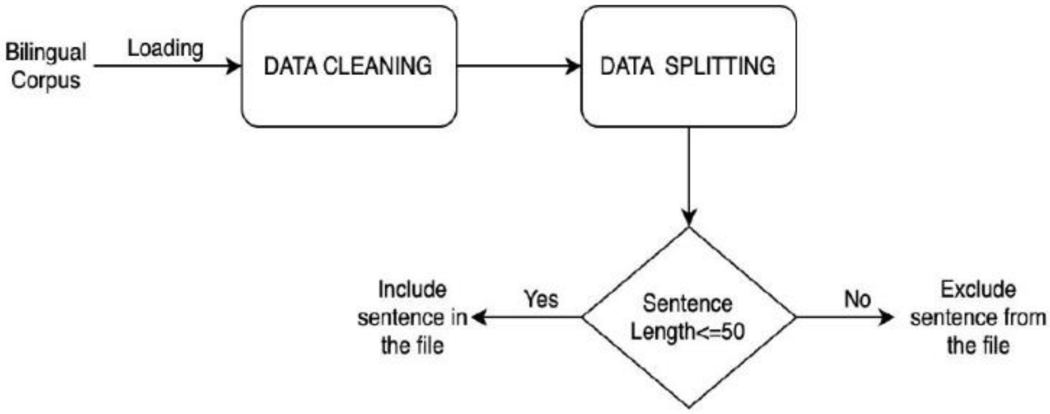

Figure 3. Preprocessing Steps for English German Dataset Training Time

The total time taken by the batches per epoch is tabulated below. Each epoch has 6 batches of size 100 to be trained and the sum of all the time of batches per epoch is calculated. The GPU with RAM of 25 GB is used to run the model and the time taken per epoch is given in seconds (sec). The training time for every 5 epochs increases by a slight margin. From the training time for 50000 training sentences the time for every epoch increases slightly and the average time taken for every epoch is almost 45 seconds. To minimize the time per epoch, a faster GPU with increased RAM size is needed.

3. Performance Analysis

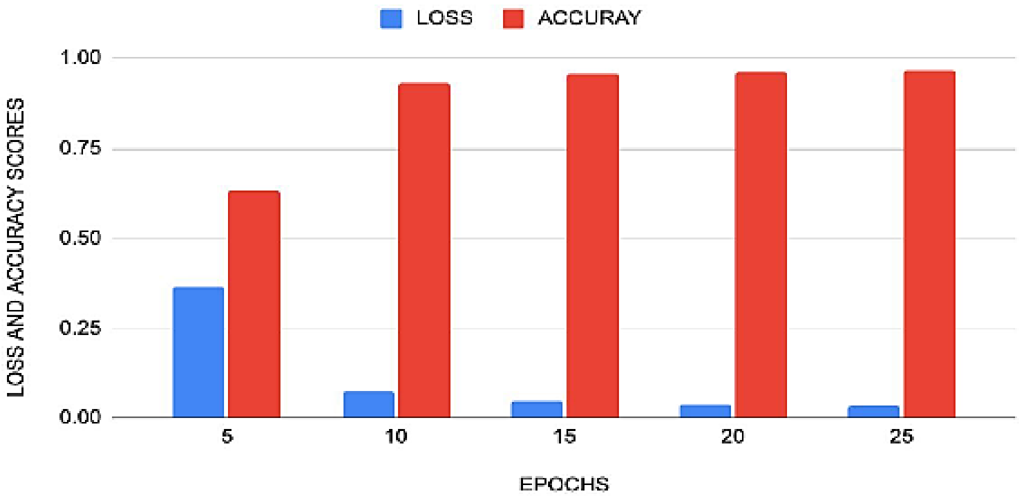

Figure 4. Bar Plot for Loss and Accuracy with Epochs

Table 1. Machine translation test cases

Test Cases	Input (German)	Actual Output	Expected Output	Pass/Fail	BLEU Score
T1	mir gehts gut.	I'm good.	i m fine.	Pass	0.93
T2	folge mir.	Follow me.	follow me.	Pass	0.99
T3	das scheint der plan zu sein.	That seems to be the plan.	it looks like half.	Fail	0.18
T4	ich habe gewonnen!	I have won!	i won!	Pass	0.85

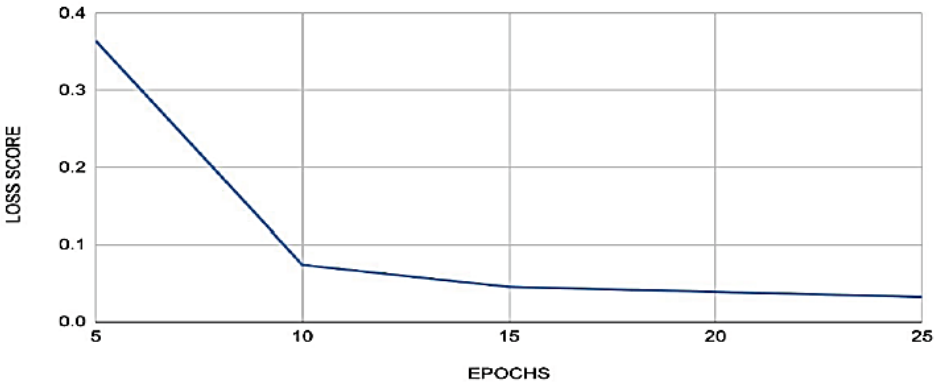

Figure 5. Graph for Training Loss Score Vs Epochs

The performance of the system is assessed based on the BLEU score and the TER score. The aim is to maximize the BLEU score so that the predicted translation is closely related to the reference translation (human interference) and to minimize the TER score so that the error in the translation becomes less. For sample 5 test cases are evaluated using the above two scores and this can be extended to the entire corpus if required.

4. Conclusion

In this work, building an end to end Language Translation system which uses a Transformer based NMT to convert German Language sentences to English Language sentences. The context of a sentence is taken into account so that the target sentence translation will be of the same meaning of the source sentence. The Self Attention mechanism used in the transformer predicts a better translated sentence. The rare words which are OOV in the sentence are handed using FSR with the help of hierarchical clustering which is integrated with the transformer model. The context of the sentence acts as the LTR which uses CNN is also integrated within the system. The Sequence to Sequence system for the dataset takes a significant time for modelling the system. The experiments indicated that the proposed NMT system achieves good performance with a better BLEU score compared to conventional MT systems and minimizes the TER significantly.

Declarations

Source of Funding

This research work did not receive any grant from funding agencies in the public or not-for-profit sectors.

Competing Interests Statement

The authors declare the total absence of conflicts of interest, both during the conduct of the research and during the written drafting of this work.

Consent for Publication

The authors declare that they consented to the publication of this research work.

References

[1] Samant, Rahul Manohar, Mrinal R. Bachute, Shilpa Gite, and Ketan Kotecha (2022). Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review and future directions. IEEE Access, 10: 17078-17097.

[2] Tonja, Atnafu Lambebo, Olga Kolesnikova, Muhammad Arif, Alexander Gelbukh, and Grigori Sidorov (2022). Improving neural machine translation for low resource languages using mixed training: The case of ethiopian languages. In Advances in Computational Intelligence: 21st Mexican International Conference on Artificial Intelligence, MICAI 2022, Monterrey, Mexico, October 24–29, 2022, Proceedings, Part II, pp. 30-40, Cham: Springer Nature, Switzerland.

[3] Bharathi, S., & T. Ananth Kumar (2020). Translation its results and insinuation in language learning. PalArch's Journal of Archaeology of Egypt/Egyptology, 17(9): 5081-5090.

[4] Heafield, Kenneth, Elaine Farrow, Jelmer Van der Linde, Gema Ramírez-Sánchez, and Dion Wiggins (2022). The EuroPat Corpus: A Parallel Corpus of European Patent Data. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Pages 732-740.

[5] Baniata, Laith H., Sangwoo Kang, and Isaac KE Ampomah (2022). A Reverse Positional Encoding Multi-Head Attention-Based Neural Machine Translation Model for Arabic Dialects. Mathematics, 10(19): 3666.

[6] Venugopalan, Manju, and Deepa Gupta (2022). An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis. Knowledge-based systems, 246: 108668.

[7] Kumar, K. Suresh, A. S. Radhamani, and T. Ananth Kumar (2022). Sentiment lexicon for cross-domain adaptation with multi-domain dataset in Indian languages enhanced with BERT classification model. Journal of Intelligent & Fuzzy Systems, Preprint, Pages 1-18.

[8] Chitnis, Rohan, and John DeNero (2015). Variable-length word encodings for neural translation models. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Pages 2088-2093.

[9] Uddin, Farid, Yibo Chen, Zuping Zhang, and Xin Huang (2022). Corpus Statistics Empowered Document Classification. Electronics, 11(14): 2168.

[10] Sun, Haipeng, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao (2020). Unsupervised neural machine translation with cross-lingual language representation agreement. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28: 1170-1182.

[11] Geng, Xinwei, Longyue Wang, Xing Wang, Mingtao Yang, Xiaocheng Feng, Bing Qin, and Zhaopeng Tu (2022). Learning to refine source representations for neural machine translation. International Journal of Machine Learning and Cybernetics, 13(8): 2199-2212.

[12] Lin, Yan, Huaiyu Wan, Shengnan Guo, and Youfang Lin (2021). Pre-training context and time aware location embeddings from spatial-temporal trajectories for user next location prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(5): 4241-4248.